A Proposed Model for Stock Price Prediction Based on Financial News

Mubarek Selimi
South East European University, Republic of North Macedonia
Adrian Besimi
South East European University, Republic of North Macedonia

Abstract

In this paper we will propose a model and needed steps that one should undertake in order to try and predict potential stock price fluctuation solely based on financial news from relevant sources. The paper will start with providing background information on the problem and text mining in general, furthermore supporting the idea with relevant research papers needed to focus on the problem we are researching. Our model relies on existing text-mining techniques used for sentiment analysis, combined with historical data from relevant news sources as well as stock data.

Keywords: text mining, finance, news, crawling, stock, prices, prediction, naïve bayes
JEL classification: C89

Introduction

The data produced and the speed at which data is produced on the Internet nowadays has increased to a degree and at a rate that is impossible to process. As such, it has encouraged research in many areas that include data mining and text mining research. These two areas have emerged in the last decade mainly due to research in artificial intelligence, machine learning, and inferential statistics (Vale, 2018).

Stock market data and relevant news associated with fin-tech industry are increasing rapidly. Many investors that are handling stock market transactions they have a major interest in understanding more about the future of stock markets for the purpose of being able to do an educated guess and/or predict any future investment. Ensuring some level of prediction in market fluctuation can assist investors in a form of decision support system and integrated with existing automatic trader agents ensure better prediction on future trades. Fully predicting the market fluctuation means in practice becoming a billionaire over the night and all the time minimizing financial losses, which is not possible for many reasons. Recent scholars argue that also news articles are influential sources that may affect stock market prices and they should be carefully considered by investors when planning future investments. By definition, any stock price is simply defined by supply and demand of the market, but it is argued by scholars Nikfarjam et al. (2010) and Kaya et al. (2010) that another important variable when decision is made to invest or not, is also related to verifiable news from financial news sources. This itself is hard and time-consuming task because it requires to read and analyse a lot of news published on several occasion by various news sources/providers (Nikfarjam, Emadzadeh and Muthaiyah, 2010; Kaya and Karsligil, 2010).

Information published in news articles influence, in a varying degree, the decision of the stock traders, especially if the given information is unexpected. It is important to analyse this information as fast as possible, so it can be used as an advantage to help traders to make trading decisions before the market has had time to adjust itself to the new information (Aase, 2011).

One important application of using text mining is text sentiment analysis, also known as opinion mining, a technique that digs deep into the content of the text file and extracts the sentiment of it. Sentiment analysis classifies textual data into positive texts, negative text and neutral text sentiments which is later used for the purpose of categorizing any text documents into the given sentiment (Aase, 2011; Khedr et al., 2017).

The model proposed in this paper is going to leverage the Naïve Bayesian classifier for document classification to make prediction for whether the stocks will go up or down, based on the dataset that is generated from the steps proposed later in this paper.

Literature Review

Several scholars, specifically Kim et al. (2014) prove somehow in their work that the relevant news are closely related to stock prices movements in the market. The current trends in big data and content creationg on the Internet and the enormous amount of unstructured text data available out there, the mobile channels, and Social Network services, scholars have attempted to predict stock movements using such text data as in the case of Kim, Jeong and Ghani in 2014.

Many scholar tried many approaches in research to prove that there is a potentially strong correlation among financial news articles and stock price fluctuations, as is the case of Khedr et al. (2017) that we mentioned earlier, in their paper they propose an approach to use sentiment analysis in financial news, along with features extracted from historical stock prices to apply prediction for the future behaviour of stocks. According to their findings, the proposed model has achieved high accuracy using sentiment analysis in categorizing news polarities by applying Naïve Bayes algorithm. In their case the accuracy of the model is up to 86.21%. By moving on with their experiment in prediction, during their next attempt in analysing these news articles, they have included also numerical attributes which in their case increased the accuracy to 89.80%.

The paper published by Hagenau et al. (2013) examines the hypothesis if any stock price prediction based on textual content from the financial news can be further improved. In this paper, the authors have upgraded the text mining methods by adding expressive feature to represent the text and by adding more variables, such as employing market feedback in the feature selection process. According to the authors, this selection of the features does significantly improve the accuracy due to the fact that this approach removes the unnecessary so called "less-explanatory features", i.e., noise, which itself helps the classifier to overcome the over-fitting during classification of the text. In the case when the feedback-based feature selection is combined with 2-word combinations, the authors results show an accuracy of 76%. These results are different from common sentiment analysis approaches since the 2-words combination give more information and potentially more meaning to the sentiment classification.

A lot of research has been carried by scholars in the area of prediction of stocks as well. A project by Joshi et al. (2016) is taking financial news articles about a given company, and they use these data to try to predict the future movements of the stock again by applying sentiment analysis. The approach is like in the other cases with an idea to identify how stocks are moved if news have polarity. Authors in this case have taken the past three years of data from Apple Inc. stock prices as well as news articles.

Similarly to previous scholars, the polarity of the news is labelling these articles and based on these data they are building the training set. The approach in this paper is dictionary-based that contains for positive and negative words that is build based on financially specific words. Futher, they have pre-processing the data which resulted in having their own finance specific stop words and dictionary. Using their own dictionary, they have implemented three models for classification and tested them. After comparing the results, they have concluded that Random Forest algorithm resulted in better accuracy for the test cases ranging from 88% to 92%. This algorithm was followed by Support Vector Machines with again very good accuracy of 86%. In their case the Naïve Bayes algorithm performance was the lowest with 83%.

Another similar approach of finding the correlation amongst the content of news articles and stock prices for the purpose of predicting the stock markets was implemented by Kaya et al. (2010). They collected news articles published in the last year period and combined with the stock prices for same period. These articles were then labelled as positive or negative sentiments categorization based on their effects on stock price. Their approach is a little bit different in the sense that for them it was important the price changes to use for categorization of the news. While analysing the textual data, authors use and approach of word doubles of a noun and a verb as features and not only single words. The support vector machines (SVM) method was used in this case which resulted with 61% accuracy.

These scholars and articles mentioned in the section above are the core of our model and study that we conducted.

Methodology

Problem definition

Financial analyst that are handling investments and transactions in stock markets around the world have a huge headache on making decision that will be effective and bring more money to the investor, or maximizing profits by trading. They are aware that any news, either good or bad can affect directly the stock market. The job of these experts relies on analysing everything from the media outlet. This is time consuming and the amount of data is getting larger and larger all the time. The methodology that we are arguing and many other scholars mentioned above as well as Falinouss (2017) is that an advanced text mining algorithms can assist these experts and provide them with knowledge just by processing resources related to text and news (Falinouss, 2007).

The price movements from the past are not always a good indicator on the future movements and are not a guarantee of smart investment, which makes news articles analysis a better predictor on stock market movements. Falinouss in 2007 proposed to research about the impact of textual data in predicting the financial markets movements. In his study he also developed a system which uses similar approaches as previous case of text mining techniques and their influence on the stock market. This according to Falinouss (2007) can help financial analysts to act immediately upon new news articles as they get published.

We propose a system of predicting stock price fluctuations or movements by analysing financial news articles on one hand and historical stock prices on the other hand. To accomplish this objective, a complete process of data mining and text mining was developed to predict the price movements for the 3 companies listed public, which are explained in the subsection bellow.

Proposed Model for Stock predictions based on financial news

In our study we worked towards analysing data, concretely news articles and historical stock prices to make future prediction about stock direction. To achieve this, qualitative and quantitative data are crucial. Many steps are conducted to achieve the aim of this research, starting from data gathering. The data is collected for a period of one year, starting from 1st of March 2018 until 1st of March 2019. In order to make the prediction we used different variables, such as the polarity of the news (either positive or negative), the rate of change in stocks quotes (an average of 5 days), a source of the news article as well as the company name.

The following are the steps needed to undertake to perform stock price prediction of financial news:

1. Identifying the news sources and targeted companies
2. Data collection and data cleaning of news articles
3. Sentiment Analysis of news articles
4. Data collection of stock prices
5. Calculating Rate of Change (ROC)
6. Categorizing the data
7. Applying Naive Bayesian classifier
8. Training

1. Identifying the news sources and targeted companies is crucial to understand your data. The information collected must be relevant and trustworthy. As such, the relevant data from financial news articles from top reliable sources have been identified as: *The Washington Post*, *CNN*, *MarketWatch*, *BGR*, *Fox Business*, *The Street*, *The Verge* and *Breitbart*. The targeted companies for our study are: *Tesla*, *Facebook* and *Apple*. News sources are proven to be reliable in the market as the most unbiased, whether the targeted companies are chosen randomly from technology, software and automotive industry. Tesla has been added because it is a typical example of a lot of news noise and several fluctuations of stock prices.

2. Data collection and data cleaning of news articles

Links of the news from the sources mentioned in step 1 are collected using Web Scraper extension of Google Chrome browser. After having all the links, we built a python script based on Scrapy[1] framework that is extracting data from the links and organizing them in the following structure: article's *Title*, *date*, *author* and the *text content*. Appropriate data cleansing has been applied to remove unnecessary HTML tags as well as to format the data from different sources to one standard (see Table 1 and Table 2).

3. Sentiment Analysis of news articles was applied to every news record based on the news content by using Vader Sentiment Analysis. VADER (Valence Aware Dictionary for sEntiment Reasoning) is a pre-built sentiment analysis model included in the NLTK package. It can give both positive/negative (polarity) as well as the strength of the emotion (intensity) of a text. VADER however is focused on social media and short texts, unlike Financial News which are almost the opposite. We updated the VADER lexicon with words plus sentiments from other sources/lexicons such as the Loughran-McDonald Financial Sentiment Word Lists, to be appropriate for our collected financial news (Yip, 2018). At the end of this step we had the polarity of the news content recorded in our dataset.

[1] Scrapy - An open source and collaborative framework for extracting the data you need from websites. www.scrapy.org

4. Data collection of stock prices for each of targeted companies was done from Yahoo Finance portal, where the following information was collected: *date*, *open price*, *high*, *low*, *close price*, *volume* and *Adj close*. These data are important for correlation with the appropriate news from our first data set.

5. Calculating Rate of Change (ROC)

The ROC and Future ROC are the two variables that are calculated from the data set from Step 4. The rate of change (ROC) in stocks in average of 5 days is an existing formula that refers to the last 5 days of stock fluctuation. In our case we also added a column with the Future ROC (the ROC after 5 days), having in mind that the effect of this positive or negative news will be reflected in the future and not the past. Since we are dealing with historical data, the Future ROC is easy to calculate.

6. Categorizing the data must be done in order to apply Naive Bayesian classifier. In the data set that we have all news collected with their features, we added two new columns: *Sentimentof_text* that could be "positive" if the sentiment score is greater than zero and "negative" if the sentiment score is less than zero. We don't take in consideration the neutral score of the text content because that could result in majority of neutral results. The second column is the *ROC_Sentiment* that can be "positive" if the *Future ROC* is greater than zero and "negative" if the the future ROC is less than zero.

7. Applying Naive Bayesian classifier

Naive Bayesian classifier was used to make the prediction of the future stock movements. The naive Bayes applies the well known Bayes' theorem, where by using a "naive" assumption that any set of features are independent for a given class (Tang, Bo; Kay, Steven; He, Haibo;, 2016). To prepare the data set to make prediction with the NB, we added a new column with the name *class* that is "UP" if the *Sentimentof_text* is "positive" and the *ROC_Sentiment* is "positive", and if the *Sentimentof_text* is "negative" and the *ROC_Sentiment* is "negative" then the class is "DOWN", otherwise is "NEUTRAL" classification. The training dataset results are summarized in Table 3.

8. Training

The data that is collected (see Table 1 and Table 2) contains records for 12 months from which 10 months will be used to train the model and the last 2 months will be used for the test set, to evaluate how it performs. In total 18236 records will be used as training dataset and the remaining 1990 records (roughly 10%) out of 20226 will be used as test set.

We created 2 models to see how they perform. The following variables are used to train and test the first model: *Source, Company, Sentimentof_text* and the *5-day ROC* while in the second model only these variables: *Source, Company and Sentimentof_text*.

Results

As explained in the steps undertaken to perform our prediction, the data collection results are shown below. We succeeded to collect the news articles from 8 different news sources, totalling 20226 news articles, split into Table 1 for Training Set (18236 records/articles) and Table 2 for Test Set (1990 records/articles).

Table 1
Total News Articles Obtained for Apple, Tesla and Facebook Organized by Source For Training Set. Period March 2018-December 2018

Variable	Categories	Frequencies	%
Source	BGR	1073	5.884
	Breitbart	435	2.385
	CNN	687	3.767
	Fox Business	813	4.458
	The Street	3810	20.893
	The Verge	2847	15.612
	The Washington post	6051	33.182
	market-watch	2520	13.819
Company	**Apple**	7591	41.626
	Facebook	7513	41.199
	Tesla	3132	17.175

Source: Authors' work

Table 2
Total News Articles Obtained for Apple, Tesla and Facebook Organized by Source for Test Set. Period January 2019-March 2019

Variable	Categories	Frequencies	%
Source	BGR	185	9.30
	Breitbart	167	8.39
	CNN	211	10.60
	Fox Business	147	7.39
	The Street	590	29.65
	The Verge	603	30.30
	The Washington post	87	4.37
	market-watch	0	0
Company	**Apple**	1144	57.49
	Facebook	416	20.90
	Tesla	430	21.61

Source: Authors' work

The training dataset results are summarized in Table 3, where for each company in our target list the classification results are shown. As general finding is that the algorithm applied classifies 15.71% of the articles in the training set as "DOWN" (meaning the stock will go down in the following days), 50.71% is classified as "NEUTRAL" (there is no clear picture on what the prediction will be) and 33.59% of the data as "UP" (meaning the stock will go up). The "UP" classification is relevant to our study and can be used for simulating investments on our test data from the test set.

Table 3
Training Set Classification Data Organized by Company and Frequency

	CLASS			
Company	**DOWN**	**NEUTRAL**	**UP**	Total
Apple	1,006	3,930	2,655	7,591
%	*5.52*	*21.55*	*14.56*	*41.63*
Facebook	1,390	3,683	2,440	7,513
%	*7.62*	*20.20*	*13.38*	*41.20*
Tesla	468	1,634	1,030	3,132
%	*2.57*	*8.96*	*5.65*	*17.17*
Total	2,864	9,247	6,125	18,236
%	*15.71*	*50.71*	*33.59*	*100.00*

Source: Authors' work

In the first prediction that uses the following variables: *Source, Company, Sentimentof_text* and the *5-day ROC* model, the test set classification from 1900 records being tested, resulted in 564 "down" and 1426 "up" classes for stock price direction were predicted.

To test the second model 3 variables as input are given: Source, Company and Sentimentof_text to predict the class up, down or neutral. comparison with the first model that has an accuracy of 94.29% the second model has 49.49% which is significantly lower accuracy than the first model that has just one more variable the 5-day ROC. It can be stated that aside from sentiment of text, stock data variables as in this case "5-day ROC" plays a vital role in prediction of the future stock price movements.

Conclusion

The trading of stock in public companies is an important part of the economy, so in this study stocks will be analysed through using data mining and text mining techniques to make a prediction for stock price directions of the stocks for 3 companies listed public.

To achieve a prediction we gathered data, collected relevant financial news articles from reliable sources both qualitative and quantitative data. This combined with the second type of data of stock prices were used in our study. For every article, a sentiment score (positive and negative) of the text content is calculated.

We have found out that a model that does not include price fluctuations and wholly relies on text content to predict the stock price fluctuation is not accurate at all. Including additional variables improves significantly the prediction. In our case the variable "5-day ROC" plays a vital role in predicting the future stock prices.

This paper is limited only on the model used and the steps undertaken to arrive at the desired data set that can be used for further prediction. Following this paper, a detailed article with results will be send for publication.

References

1. Aase, K. G. (2011), "Text Mining of News Articles for stock Price Prediction", Master's thesis, Institutt for datateknikk og informasjonsvitenskap.
2. Falinouss, P. (2007), "Stock Trend Preidction Using News Articles: A Text Mining Approach", Master's Thesis.

3. Hagenau, M., Liebmann, M. Neumann, D. (2013), "Automated news reading: Stock prices prediction based on financial news using context-capturing features", Decision Support Systems, Vol. 55, No. 3, pp. 685-697.
4. Joshi, K., Rao, J., Bharathi, H. N. (2016), "Stock Trend Prediction Using News Sentiment Analysis", International Journal of Computer Science & Information Technology (IJCSIT), Vol. 8, No. 3, pp. 67-76.
5. Kaya, M. Y., Karsligil, M. E. (2010), "Stock price prediction using financial news articles", in the Proceedings of the 2nd International Conference on Information and Financial Engineering, Chongqing, China, IEEE, pp. 478-482.
6. Khedr, A. E., Yaseen, N. (2017), "Predicting stock market behavior using data mining technique and news sentiment analysis", International Journal of Intelligent Systems and Applications, Vol. 9, No. 7, pp. 22-30.
7. Kim, Y., Jeong, S. R., Ghani, I. (2014), "Text opinion mining to analyze news for stock market prediction", International Journal of Advances in Soft Computing and its Application, Vol. 6, No. 1.
8. Nikfarjam, A., Emadzadeh, E., Muthaiyah, S. (2010), "Text mining approaches for stock market prediction", in the Proceedings of the 2nd International Conference on Computer and Automation Engineering (ICCAE), Singapore, Singapore, IEEE, pp. 256-260.
9. Vale, M. N. d. (2018), "Dow Jones Index Change Prediction Using Text Mining", Instituto Alberto Luiz Coimbra De Pós-Graduação E Pesquisa De Engenharia, Rio de Janeiro, Brazil.
10. Yip, J. (2018), "Algorithmic Trading using Sentiment Analysis on News Articles", available at: https://towardsdatascience.com/https-towardsdatascience-com-algorithmic-trading-using-sentiment-analysis-on-news-articles-83db77966704 (15 May 2019).

About the authors

Mubarek Selimi has a Bachelor in Business Informatics and currently pursuing his Master of Science degree in Business Informatics. His interest mainly are: Business Information Systems, E-commerce, Data Mining, Business Intelligence and Financial analysis. The author can be contacted at **ms21693@seeu.edu.mk.**

Adrian Besimi has a PhD in computer science and is employed as an Associate Professor at the CST department of SEE University. His interests are mainly related with application of IT in business and other organizations, such as: Business Information Systems, e-Commerce, Web and Mobile Solutions, Software Oriented Architectures, Data Mining and Business Intelligence and similar. The author can be contacted at **a.besimi@seeu.edu.mk.**